政治学方法論 I 課題 6

提出者：宋財泫 (123J009J)

提出日：2014-11-09(日)

1　問題 1

重回帰分析のシミュレーションを行い、以下の各問いに答えなさい (R コードの提出は不要)。母数 (parameters)、サンプルサイズ、説明変数の取り方等は自分で設定すること。ただし、すべての問題について、複数のサンプルサイズを比較検討すること。すべての内容が含まれていれば、回答の順番は変えてもよい。

1.1　シミュレーション

1.1.1　共通

重回帰分析シミュレーションのために関数 Jay.ols を作成する。

```
Jay.ols <- function(beta, sigma, n=100, trials=10000, x.range
    =c(0,10), ci.level=0.95){
   empty <- rep(NA, trials)

   df <- data.frame(
    id = empty,
    b1 = empty,
    b1.se = empty,
    b1.lower = empty,
    b1.upper = empty,
    b2 = empty,
    b2.se = empty,
    b2.lower = empty,
    b2.upper = empty,
    b3 = empty,
    b3.se = empty,
```

```
    b3.lower = empty,
    b3.upper = empty,
    check.ci = empty,
    check.ci2 = empty,
    sigma.hat = empty)

  for(i in 1:trials){
    x1 <- runif(n, x.range[1], x.range[2])
    x2 <- runif(n, x.range[1], x.range[2])
    y <- rnorm(n, beta[1] + beta[2] * x1 + beta[3] * x2,
        sigma)

    fit <- lm(y ~ x1 + x2)

    sigma.hat <- summary(fit)$sigma

    b1 <- coef(fit)[1]
    b2 <- coef(fit)[2]
    b3 <- coef(fit)[3]

    b1.se <- sqrt(summary(fit)$cov.unscaled[1, 1]) * sigma.
        hat
    b2.se <- sqrt(summary(fit)$cov.unscaled[2, 2]) * sigma.
        hat
    b3.se <- sqrt(summary(fit)$cov.unscaled[3, 3]) * sigma.
        hat

    b1.ci <- confint(fit, level=ci.level)[1, ]
    b2.ci <- confint(fit, level=ci.level)[2, ]
    b3.ci <- confint(fit, level=ci.level)[3, ]

    check.ci <- b2.se[1]<=beta[2] & b2.se[2]>=beta[2]
    check.ci2 <- b3.se[1]<=beta[3] & b3.se[2]>=beta[3]

    df[i, ] <- c(i, b1, b1.se, b1.ci,
                 b2, b2.se, b2.ci,
                 b3, b3.se, b3.ci,
                 check.ci, check.ci2, sigma.hat)
  }

  return(df)
}
```

1.1.2　回帰係数の性質 (問題 1–1)

統計量やヒストグラムなどから、回帰係数の性質を調べなさい。

切片が 5, 説明変数の係数が 10 と 3, 標準偏差が 8 のモデルを 10,000 回施行し、ヒストグラムと統計量を確認する。

```
Jay.sim <- Jay.ols(beta=c(5, 10, 3), sigma=8, trials=10000)
```

β_2 の統計量は以下のようである[1)]。表 1 から β_2 の平均が 10 であり、標準偏差も非常

表 1　β_2 の記述統計

平均	最小値	最大値	標準偏差	歪度	尖度
10.001	9.079	11.021	0.283	0.008	3.001

に小さいことが分かる。また歪度と尖度が正規分布 (歪度=0, 尖度=3) に極めて近似していることから β_2 が正規分布の形をしていることが予想される。したがって、ヒストグラムと Q-Q プロットでそれを確かめる。

次のページの図 1 と 2 は β_2 のヒストグラムと Q-Q プロットであり、記述統計量で見たように正規分布に近い形をしていることが確認できる。Q-Q プロットでは理論的な正規分布[2)]からやや外れた観測値が一部観察できるが、概ね正規分布と言っても差し支えないとも言えよう。

続いて係数の標準誤差の意味であるが、これは複数回の試行から得られた係数の標準偏差として理解する事ができる。実際に確かめてみると β_2 の標準偏差は約 0.28274 であり、$S.E._{\beta_2}$ の平均値は約 0.27851 である[3)]。この差は約 0.00423 であり、ほぼ一致することが分かる。これは試行回数を増やすとより差が縮まると予想される。

---

1) 切片や β_3 に対して同様の手順で行っても得られる結果は変わらないため、β_2 のみについて述べる。

2) この場合は表 1 から平均 10, 標準偏差 0.28285 の分布を想定する

3) $\sigma = 8, N = 10,000$ の場合、理論的には $\sigma_{\beta_2} = \overline{S.E._{\beta_2}} = \sqrt{\frac{8}{\sqrt{100}}} \approx 0.28284$

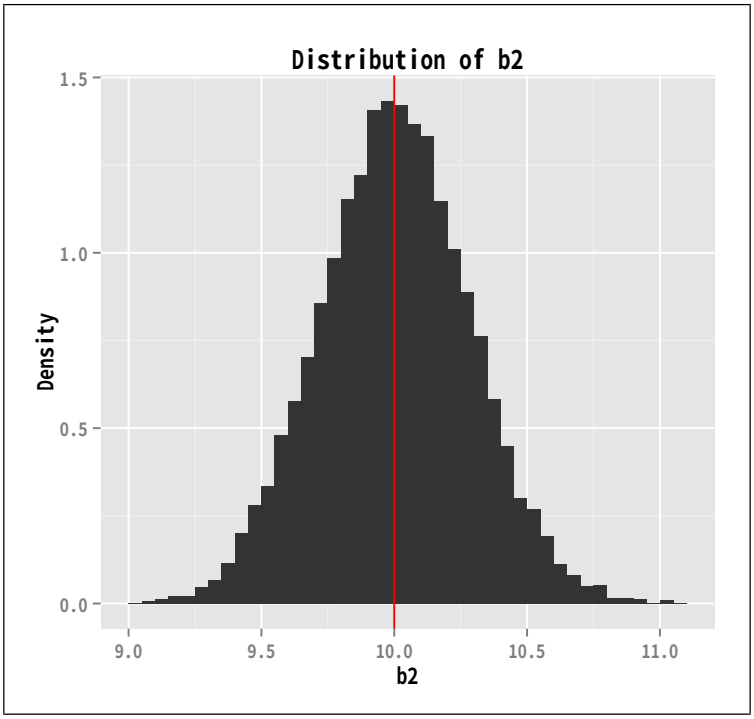

図 1　β_2 の分布

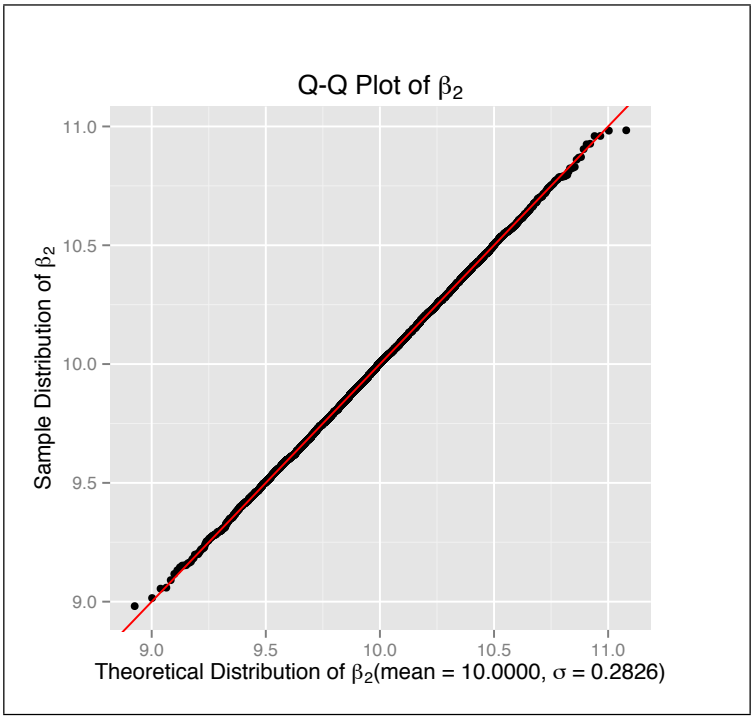

図 2　β_2 の Q-Q プロット

1.1.3　95% 信頼区間の性質 (問題 1–2)

95 パーセント信頼区間の性質を調べなさい。その際、シミュレーションから得られた 100 個の 95 パーセント信頼区間を 1 つの図に示しなさい。

```
sim <- Jay.ols(beta=c(5, 6, 3), sigma=4, trials=100)

sim.result <- ggplot(sim, aes(x=reorder(id, b2), y=b2,
                          ymin=b2.lower, ymax=b2.upper,
                          color=check.ci)) +
  geom_hline(yintercept=5, linetype="dotted") +
  geom_pointrange() + guides(colour=FALSE) +
  labs(x="Simulation ID", y="b2", title="95% CI") +
  coord_flip()

print(sim.result)
```

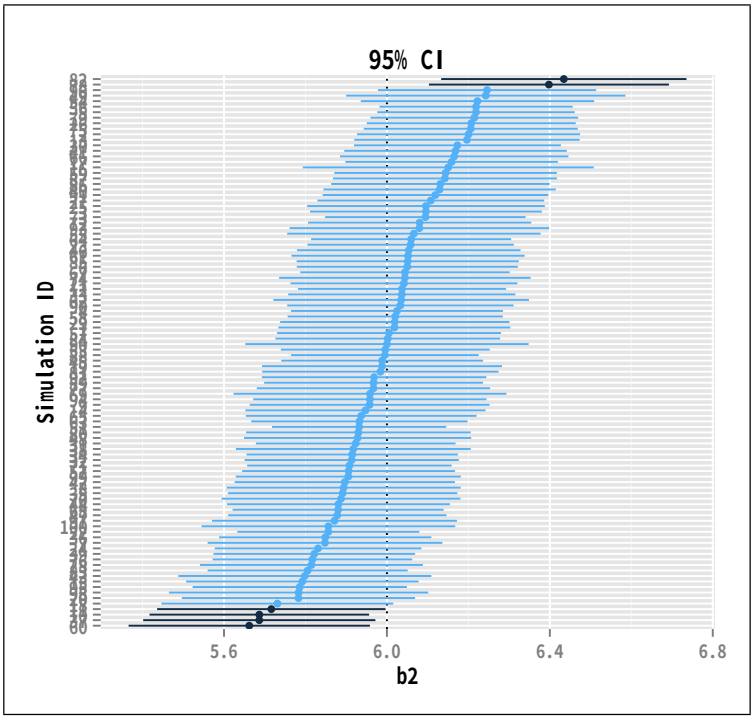

図 3　95% 信頼区間のシミュレーション

図 3 はシミュレーションの結果を表す。6 回の試行において信頼区間が母数を含まないことが確認できる (sum(sim$check.ci) で確認)。シミュレーションであるため、正確だとは言えないが、95% に近似していると言えよう。

1.1.4　50% 信頼区間の性質 (問題 1–3)

50 パーセント信頼区間の性質を調べなさい。その際、シミュレーションから得られた 100 個の 50 パーセント信頼区間を 1 つの図に示しなさい。

```
sim2 <- Jay.ols(beta=c(3, 8, 6), sigma=9, trials=100, ci.
    level=0.50)

sim.result2 <- ggplot(sim2, aes(x=reorder(id, b2), y=b2,
                          ymin=b2.lower, ymax=b2.upper,
                          color=check.ci)) +
  geom_hline(yintercept=8, linetype="dotted") +
  geom_pointrange() + guides(colour=FALSE) +
  labs(x="Simulation␣ID", y="b2", title="95%␣CI") +
  coord_flip()

print(sim.result2)
```

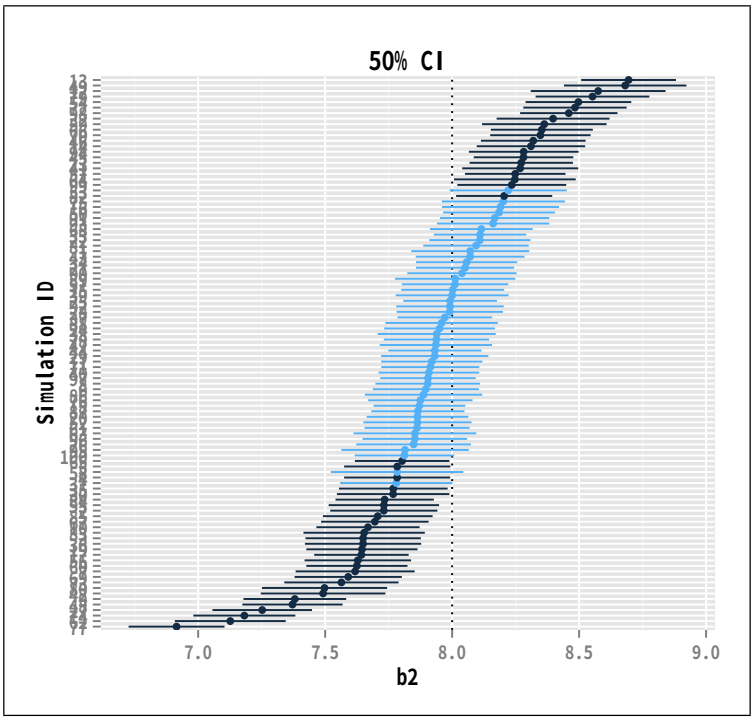

図 4　50% 信頼区間のシミュレーション

図 4 はシミュレーションの結果を表す。50 回の試行において信頼区間が母数を含まないことが確認できる (sum(sim2$check.ci) で確認)。シミュレーションであるため、偶然

とも言えるが、ちょうど 50 回において信頼区間にパラメータを含んだ結果が得られた。問題 1–2 の結果と合わせて考察すると x% の信頼区間とは同じ母集団から複数回の試行でサンプルを抽出し、同じモデルで推定する場合、x% の試行が信頼区間に母数を含むことを意味することが分かる。

1.2 Model Mis-Specification

1.2.1 Omitted Variables(問題 1–4)

・データ生成に使った説明変数 (つまり、実際に応答変数に影響を与えている変数) を 1 つ以上取り除いた回帰分析を行い、回帰係数と残差の性質を調べなさい。

最初に $x1$ と $x2$ を両方投入したモデル (True Model) を推定し、続いてデータ生成過程で使われた $x2$ を除いたモデル (Omitted Model) を推定して比較する。

$$TrueModel : y = X_1\beta_1 + X_2\beta_2 + \upsilon \tag{1}$$
$$OmittedModel : y = X_1\beta_1' + \varepsilon \tag{2}$$

表 2　問題 1–4 のパラメータ設定

切片	β_1	β_2	σ
5	3	8	10

式 1 の X_1 は切片と $x1$ の行列であり、X_2 は $x2$ のベクトルである。式 2 も同様であるが、式 1 と異なるのは誤差項のみである y は上記のデータ生成過程から得られた値の行列である。x_1 と x_2 は 1 から 10 の一様分布 ($U(1, 10)$) から抽出し、この過程を 100 回繰り返した結果をベクトル X_1, X_2 とする。

モデルとパラメータを設定した後、問題 1(共通) のようにシミュレーションを行う[4)]。シミュレーションの試行回数は 10,000 回である。シミュレーションから得られたパラメータの算術平均値は以下のようである。

表 3 を見ると Omitted Model の推定値が Ture Model の推定値と大きく離れていることが確認できる。これを視覚的に理解するために β_1 と β_2' のヒストグラムと残差の標準偏差 σ のヒストグラムを示す。

図 5 をみると $4\beta_1$ は母数である 3 を中心に分布しているが、β_1' は 4.6 の周辺に分布している。また、図 6 からは残差の標準偏差が True Model において小さく、最初に設定した 10 に近い。残差とはモデルによって説明されなかった部分を指すので、この場合、True Model の方が Omitted Model より y をよく説明できると言えよう。Omitted Model において説明されなかった部分とは欠落した変数 x_2 によって説明されうる部分である。

[4)] 上記の関数をそのまま使わず、変数の数に合わせて再調整した

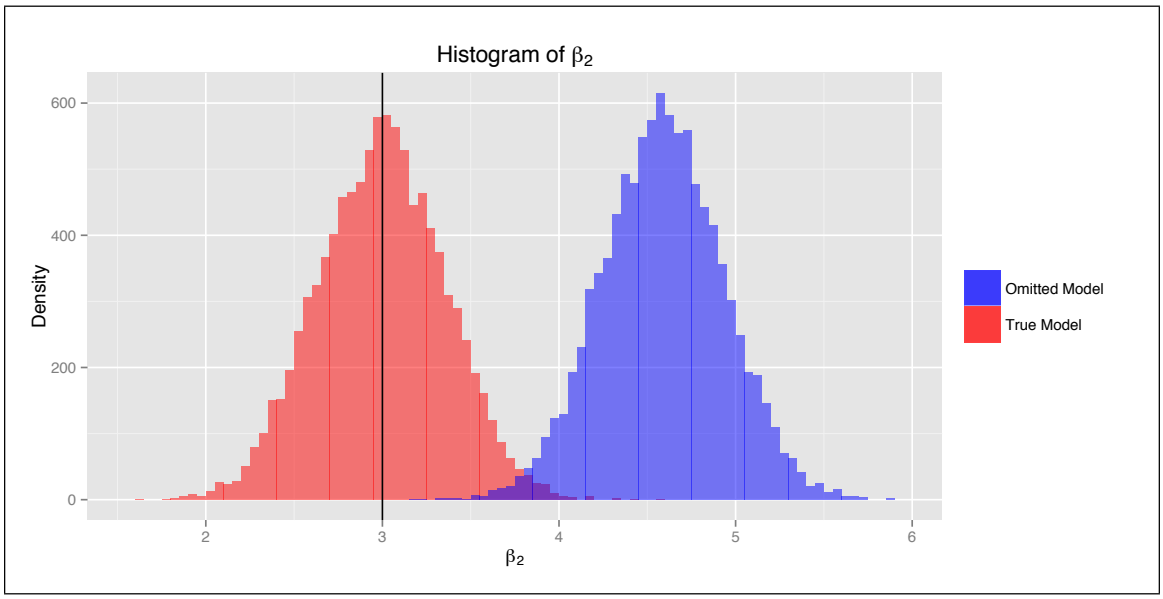

図 5　β_1 と β_1' の分布の比較

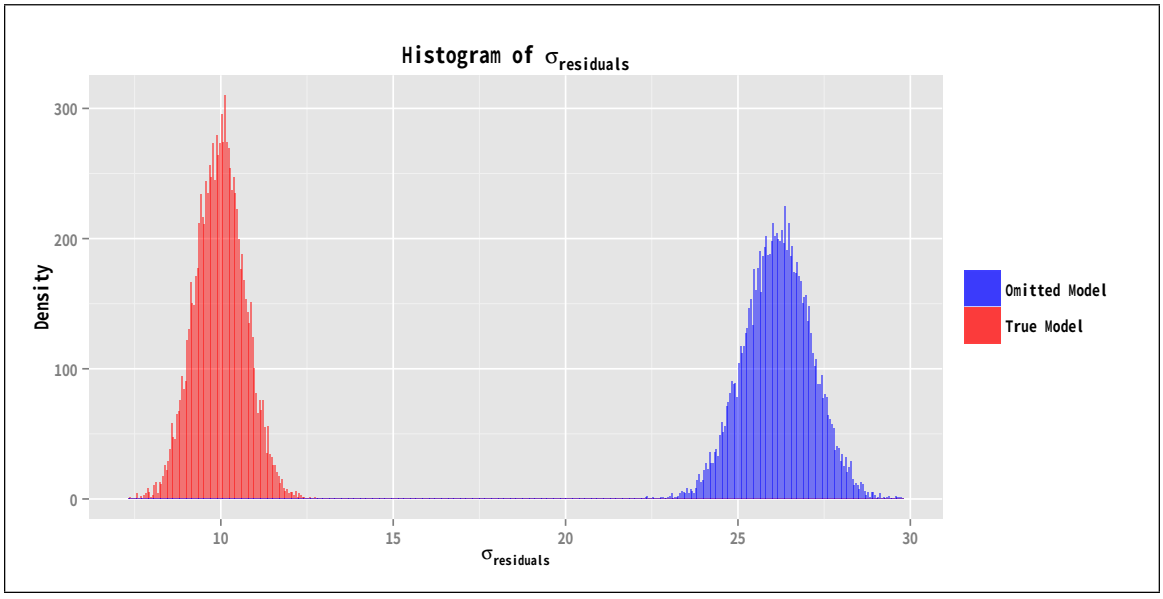

図 6　True Model と Omitted Model の σ の分布の比較

表 3　True Model と Omitted Model の推定結果の比較

	Ture Model	Omitted Model
Intercept	5.000(2.371)	36.67(5.166)
$x1$	2.995(0.354)	4.587(0.9199)
$x2$	8.005(0.333)	
σ	9.983	26.15
注：括弧内は標準誤差		

残差と欠落した変数のパラメータには関係がある。そのために残差と欠落された変数 $x2$ の回帰係数 β_2 の関係について数式的に調べる。この関係は上記の式 1 と 2 から以下のように表現できる。

$$\begin{aligned} X_1\beta_1 + X_2\beta_2 + \upsilon &= X_1\beta_1' + \varepsilon \\ X_1\beta_1 + X_2\beta_2 - X_1\beta_1' &= \varepsilon - \upsilon \\ X_1(\beta_1 - \beta_1') + X_2\beta_2 &= \varepsilon - \upsilon \\ X_2\beta_2 &= \varepsilon - \upsilon - X_1(\beta_1 - \beta_1') \end{aligned}$$

つまり、β_2 は Omitted Model で説明されなかった残差 (ε) と True Model で説明されなかった残差 (υ; 真の誤差) の差に True Model と Omitted Model において $x1$ によって説明された部分 ($X_1(\beta_1 - \beta_1')$) を引いた値を応答変数として X_2 に切片なしで回帰させた時の係数となる[5)]。また、ここでは β_1 と β_1' はかなり違うが、そのバイアスの程度は以下のように表現できる。

$$\begin{aligned} \beta_1' &= (X_1^T X_1)^{-1} X_1^T y \\ \beta_1' &= (X_1^T X_1)^{-1} X_1^T [X_1\beta_1 + X_2\beta_2 + \upsilon] \\ \beta_1' &= \underbrace{(X_1^T X_1)^{-1} X_1^T X_1}_{I}\beta_1 + (X_1^T X_1)^{-1} X_1^T X_2\beta_2 + \underbrace{(X_1^T X_1)^{-1} X_1^T \upsilon}_{E[\upsilon]=0} \\ \beta_1' &= I\beta_1 + (X_1^T X_1)^{-1} X_1^T X_2\beta_2 + 0 \\ E[\beta_1'] &= \beta_1 + E[(X_1^T X_1)^{-1} X_1^T X_2\beta_2] \end{aligned} \tag{3}$$

[5)] もし X_1 と X_2 が完全に独立している場合は $X_1(\beta_1 - \beta_!') = 0$ であり、残差の差だけで十分

式3により Omitted Model における β_1' は Ture Model の β_1 に $E[(X_1^T X_1)^{-1} X_1^T X_2 \beta_2]$ だけのバイアスが含まれている係数だということが分かる。

1.2.2 Overfitting Variables(問題 1–5)

・データ生成に使わなかった説明変数 (つまり、実際には応答変数に影響を与えない変数) を 1 つ以上加えた回帰分析を行い、回帰係数と残差の性質を調べなさい。

ここではデータ生成過程において応答変数に影響を与えなかった変数 $x2$ を用いる。$x2$ も $x1$ と同様、1 から 10 までの一様分布から抽出された変数 ($U(1,10)$) である。モデルは以下のようである。

$$TrueModel : y = X_1\beta_1 + \upsilon \tag{4}$$
$$OverfittingModel : y = X_1\beta_1' + X_2\beta_2 + \varepsilon \tag{5}$$

表 4　問題 1–5 のパラメータ設定

切片	β_1	σ
3	12	4

X_1 と X_2 は $U(1,10)$ から 100 回抽出された $x1$ と $x2$ のベクトルである。$x2$ はデータ生成過程で使われなかった変数である。このモデルに基いて試行回数 10,000 回のシミュレーションを行った。まずは、True Model と $x3$ を投入した Overfitting Model の推定値の平均値を比較する。

表 5　True Model と Overfitting Model の推定結果の比較

	Ture Model	Overfitting Model
Intercept	2.987(0.801)	2.993(1.033)
$x1$	12.000(0.134)	12.000(0.134)
$x2$		0.002(0.136)
σ	3.992	3.984
注：括弧内は標準誤差		

結果を比較すると $x2$ を投入した Overfitting Model は True Model と比べて説明され

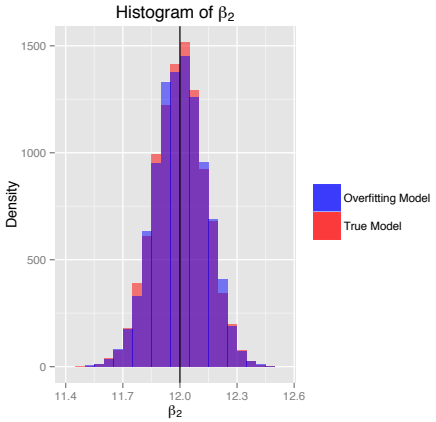

図 7　β_1 と β_1' の比較

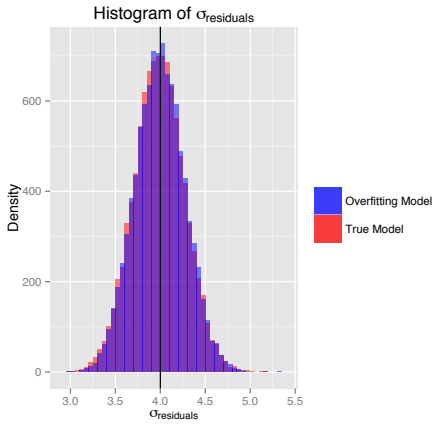

図 8　σ の比較

なかった係数と残差がほぼ変わらないことが確認できる。これを視覚的に表すために β_1, β_1' のヒストグラムと σ のヒストグラムを以下に示す。

ヒストグラムを見ると両モデルにおいて $x1$ の係数はパラメータである 12 の周辺に分布し、残差の標準偏差も 4 を中心に分布していることが確認できる。これは Overfitting Model が True Model に比べてバイアスは生じさせないものの、より良い結果を出してはいないとも言えよう。実際に、回帰係数がほぼ変わらない理由は以下のように導出することができる。説明変数が 2 つ (x, z) のみのモデルを例として考える。

$$\beta_x = \frac{Var_z Cov_{y,x} - Cov_{x,z} Cov_{y,z}}{Var_x Var_z - Cov_{x,z}^2}$$
$$\beta_x = \frac{r_{y,x} - r_{x,z} r_{y,z}}{(1 - r_{x,z}^2)} \frac{\sigma_y}{\sigma_x} \tag{6}$$

式 4 に基き、z が x と y に対して独立 (共分散、相関係数がゼロ) していると仮定する。

$$\beta_x = \frac{r_{y,x} - 0}{(1 - 0)} \frac{\sigma_y}{\sigma_x} = r_{y,x} \frac{\sigma_y}{\sigma_x} \tag{7}$$

つまり、β_x は単回帰分析における x の係数と一致する。z と他の変数の間で共分散が大きくなれば、バイアスが生じうるが、データ生成過程から独立している場合は共分散が非常に小さく、回帰係数には大きな影響を与えない事が式 5 より確認できる。

Overfitting の問題は推定値の不偏性よりはモデルの効率性 (efficiency) と関係する。モデルにおいて説明された分散の割合とパラメータの数を総合的に考慮した効率性の指標ともみなせる指標の平均値を比較すれば以下の表のようである。

表 6　True Model と Overfitting Model のモデル比較

	Ture Model	Overfitting Model
Adj. R^2	0.988	0.987
AIC	560.485	571.117
BIC	568.300	581.538

調整済み決定係数 (Adj. R^2) は大きいほうが、AIC と BIC は小さいほうがより良いモデルの基準となるが、全ての統計量において True Model の方がより効率性の高いモデルであることを示している。したがって、応答変数と説明変数に独立しているから説明変数を投入した Overfitting Model はバイアスはほぼ生じさせないものの、効率性の面で劣る事が確認できる。

2　問題２

最小二乗法による回帰分析に使えそうなデータセットを見つけなさい。ただし、応答変数となり得る変数を 1 つ以上、(各) 応答変数の説明変数となり得る変数が 2 つ以上含まれるものを見つけること。複数のデータソースを使い、自分でデータセットを作ってもかまわない。変数名をすべて英数字に変換した長方形データを csv ファイルとして提出しなさい。

別ファイルとして提出 (data4ols-SONG.csv)